Comparing Tree-Based, MLP, and CNN Models for ECG-Based Sudden Cardiac Death and Atrial Fibrillation Prediction

Roni Ahola[1], Jussi Hernesniemi[1,2], Antti Vehkaoja[1]

[1]Tampere University, Tampere, Finland
[2]Tampere Heart Hospital, Tampere, Finland

Abstract

Published ECG analysis models vary in architecture, input representation, training objective, and evaluation protocol, and attribution given to the model architecture regarding the performance may be misleading.

We used three cohorts from two hospitals to predict two endpoint events, SCD and AF, from ECGs parametrized in four different ways, with two different training objectives, and overall ten different models. We used hyperparameter optimization with nested CV and a generous computing budget to compare these.

The model architecture was shown to matter very little — 0.02–0.03 movement of cause-specific C-index in the 0.75 range — although a consistent winner was XGBoost, and consistent loser the only linear model, sparse logistic regression. In transfer learning, the direction of the transfer was important, and the raw waveform ECG transferred better than feature-extracted ECG.

We found XGBoost to be both the best and the fastest model architecture, and therefore conclude that the choice between architectures should be based on other aspects, such as the availability of secondary analysis, such as explainable AI methods.

1. Introduction

Electrocardiography (ECG) has been widely adopted by the machine learning and deep learning communities as an input data modality for cardiovascular risk modelling. Especially deep learning based on the raw waveform ECG has seen large literature adaptation [1], [2].

The models published in literature are typically compared within the same architecture family, in a tight modelling space, which makes the AUROC gap between two papers unattributable to a specific choice in methodology. It is not exactly understood whether the current favour for deep learning models stem simply from the larger interest, or a real superiority of the models [3], [4].

The complexity of an ECG may be much smaller than an image, but the underlying mechanisms may require latent representations that simpler models do not provide. Yet, the ECG is also possible to disassemble, or delineate, into well-known parts, such as to the P-wave, QRS-complex, and the T-wave. And within them, further on into features such as time of the R-peak, amplitude of the wave's features, fiducial points such as the J-point, and slopes between these points.

In our study we test whether the ECG can be utilized equally well using a tabulated feature-set as with the raw waveform. Deep learning experience has shown that almost any signal is recoverable from raw data, with the proper neural network and learning task. We use two different types of architectures to process the raw waveform. We compare these to three different types of tabular representations of the same signal, and further compare different modelling techniques within the tabular-feature architectures available.

We further evaluate the different types of predictions available for each architecture, ranging from binary, to discrete-time survival prediction, to continuous time survival prediction. Finally, we also test how transfer learning affects the ranking between the architectures, and how fast in computation time each modelling method is. Our analysis is focused on high-sized datasets.

2. Methods

2.1. Cohorts and endpoints

Our analysis included three cohorts from two datasets, with two different endpoints. MADDEC [5] is a single-site dataset from Tampere, Finland, and includes patients who were recruited based on referral to coronary angiography. It includes multiple ECGs for each patient, with median of 16-18, over median follow-up of 7.0-7.1 years. MIMIC-IV-ECG [6], [7] is a single-site dataset from the US. It has markedly fewer ECGs per patient, with median of 2 over 1.6 follow-up years. More cohort information can be found in Table 1.

Table 1. The three cohorts, median used in averages.

	MADDEC-AF	MADDEC-SCD	MIMIC-IV
ECGs	233 820	460 154	538 228
Patients	12 635	21 175	139 651
ECGs per patient	16	18	2
Age at ECG, y	70	73	62
Female, %	41.1	39.6	51.4
Event ECGs	48 626	35 350	78 298
Competing death ECGs	51 762	150 917	103 032
Follow-up, y	7.1	7.0	1.6
Known outcome ECGs at 2 y	216 797	425 567	349 523
Prevalence at 2 y, %	7.5	3.1	14.3

Our two endpoints are sudden cardiac death (SCD), and atrial fibrillation or flutter (AF). The SCD endpoint exists only in MADDEC, while the AF endpoint enables cross-cohort comparisons as well. The AF endpoint is constructed using the ECG-cart reading: the AF-onset is encoded as the date of the first ECG where AF is present, according to the ECG.

2.2. ECG data representations

The ECG signal data is represented in four different ways. The original signal data is used as a 12-lead, 250 Hz, or 100 Hz for the foundation model, and 10-second segments. Second, we get embeddings from the frozen foundation-model for MADDEC data. (Foundation model was partly pretrained with MIMIC data.) Third, we use the GE 12SL parameter set [8]: 1266 parameters for AF and 1355 for SCD, computed from the raw ECG by a proprietary algorithm. This set also only exists for MADDEC. Fourth, we have a set of features from our own internal delineator, that provides 406 features delineated from the ECG.

Only the raw waveform and the 406 delineated feature set allow for cross-cohort analysis, while MADDEC is further analyzed using the 12SL and the FM features due to data availability. We included our own delineator as this allowed for cross-cohort analysis for tabular data, because the 12SL parameters were not available for MIMIC.

2.3. Architectures and objectives

Further on, to separate the representation from the architecture, we analyzed multiple architectures for each data representation. For the raw waveform, we used two different models: a CNN model, and the foundation model with low-rank adaptation, to more robustly validate the waveform approach besides a single model. Tabular data were modelled using four different architectures: sparse logistic regression, LightGBM, MLP, and XGBoost.

Out of the six different architectures, four were evaluated under two different objectives: binary and survival prediction, under a 2-year prediction horizon. Sparse logistic regression was evaluated only under binary, and FM only under survival. The survival is discrete under MLP-hazard, CNN-hazard, and FM-hazard, and continuous under XGB-AFT and LGBM-Cox.

2.4. Evaluation

We used a nested 5-fold cross-validation scheme with patient grouping. To optimize hyperparameters for each architecture, and objective, we allowed each to have 60 Optuna [9] trials, with 2-year binary AUROC or 2-year Harrell's C-index being the Optuna objective. The last 20 trials brought +0.0006 AUROC units on average, with 0.00216 as max. FM was the exception, and it was given default parameters. Overall, 16,500 trials were held across all the sets.

We report the C-index as the primary result, with AUROC using patient-clustered bootstrap intervals as a secondary metric. Each model's best hyperparameter configuration was used to give the score in 18 different ways. First, the chosen ECGs to be used had three different methods: all ECGs available, first ECG per patient, last ECG per patient before the endpoint event. Then six label rules for each: competing deaths either kept as negatives or dropped, censored before the horizon either kept or dropped, event "ever" ignoring the horizon, and finally the most permissive setup where all competing deaths ever were dropped, and censored events within the horizon window were dropped.

For transfer learning we fit all five outer folds' chosen configurations on the full source cohort and ensembled their prediction on the whole target cohort. Tabular features were scaled using the median/IQR of the target cohort without target labels.

3. Results

3.1. Architecture within a representation

For both endpoints, the choice of architecture ended up mattering quite little. Table 2 shows that while XGBoost won six of the seven cohort × representation combinations it was run in, the swing from worst to best was only a few points in the metric. Sparse logistic regression was shown to be consistently the worst, which suggests that nonlinearity is needed to some extent.

The architecture ordering was consistent between MADDEC and MIMIC, with a spearman correlation of ρ=0.893 on the delineator features. The same ranking indicates that this classifier ranking is general to ECG in this task and size of data, rather than a cohort related property. Finally, the task between binary and survival had a clear winner: survival beat 23 of the 24 matched pairs.

3.2. Representation differences

From Table 2 we can see that the representations also seemed to matter very little. The waveform CNN is below the best 12SL models on both MADDEC endpoints. Against the delineator features it wins on SCD on all seven models, but loses to four out of seven on AF, and to five out of seven on MIMIC. The point difference was similar in MADDEC and MIMIC.

Table 2. Cause-specific Harrell C at 2 years, nested 5-fold cross-validation.

	MADDEC-AF	MADDEC-SCD	MIMIC-IV
12SL parameters (1266/1355)			
XGB-AFT	**.7591**	**.7808**	—
XGB-binary	.7534	.7728	—
LGBM-Cox	.7525	.7758	—
LGBM-binary	.7528	.7737	—
MLP-hazard	.7470	.7753	—
MLP-binary	.7467	.7613	—
Sparse logistic regression	.7287	.7509	—
Delineator features (406)			
XGB-AFT	**.7575**	**.7734**	**.7717**
XGB-binary	.7519	.7657	.7667
LGBM-Cox	.7506	.7677	.7676
LGBM-binary	.7496	.7658	.7667
MLP-hazard	.7529	.7695	.7671
MLP-binary	.7452	.7570	.7639
Sparse logistic regression	.7343	.7470	.7429
FM embedding (1024)			
XGB-AFT	**.7460**	.7733	—
XGB-binary	.7380	.7676	—
LGBM-Cox	.7417	.7694	—
LGBM-binary	.7372	.7661	—
MLP-hazard	.7430	**.7765**	—
MLP-binary	.7414	.7709	—
Sparse logistic regression	.7278	.7595	—
FM, DoRA (no search)			
FM-hazard	**.7500**	**.7739**	—
Raw waveform			
CNN-hazard	**.7504**	**.7748**	**.7648**
CNN-binary	.7437	.7536	.7634

Within the tabular format, the feature extraction methods included the foundation-model embeddings, 12SL parameters, and our own delineator parameters derived from the raw waveforms. Within these, the differences were small, even as the amount of features differed from 1024 on FM, to 406 on our delineator, to 1266 or 1355 on 12SL. It is likely that all of them represented the data well enough for the tasks.

The DoRA finetune of the foundation model differed marginally on both endpoints from the CNN, with point difference being 0.0004 on AF, and 0.0009 on SCD. For SCD, Optuna chose on median 1.97M parameters for the CNN on SCD, against 6.4M trainable parameters in the FM. CNN was searched for over 108 combinations, ranging 45.5k to 31.7M parameters, so a model from the lower end was chosen.

3.3. Effect of the negative set

The chosen set to compare against had a higher effect than any architectural or objective choice. Out of the 18 possibilities, the most lenient comparison was excluding all competing deaths, and also excluding censors within 2 years. For XGBoost binary this landed at 0.8118 AUROC, while the strictest form that uses only the last ECG per patient, drops censors, and includes competing deaths as negatives, gave AUROC of 0.6505. A sensible set for the same model gave AUROC of 0.7602 (all ECGs, competing = negative, censored dropped). Overall, the effect of architecture moved the AUROC 0.018-0.032 against a fixed negative set, while negative set moved it by 0.128-0.180 against a fixed architecture. The distance of the strictest setup to the most permissive had a larger effect than any architecture difference.

3.4. Transfer-learning

Transfer was strongly asymmetric between the two dataset sites (Figure 1). Generally, transfer learning caused a drop in C-index, but much less when trained with MIMIC and tested on MADDEC than when trained on MIMIC and tested on MADDEC. The size of the MIMIC set is larger, which can be part of the reason. Another possibility is the focus on acute ECGs (near the first event) in MIMIC versus MADDEC. If we excluded ECGs that had less than 7 days before the endpoint event, the advantage of MIMIC-transfer dampened. MIMIC is rich in acute presentations, and MADDEC is longitudinal, which might explain the advantage training with MIMIC brings to MADDEC. Out of all the transfers tested, only the CNN waveform got better, although only minimally (0.005). It is also seen here that a MIMIC-trained CNN matched the MADDEC-trained one, despite never having seen a case from there, while the same was not true for the engineered features. This might also be due to the ensemble in the transfer, as the native sets were not ensembled.

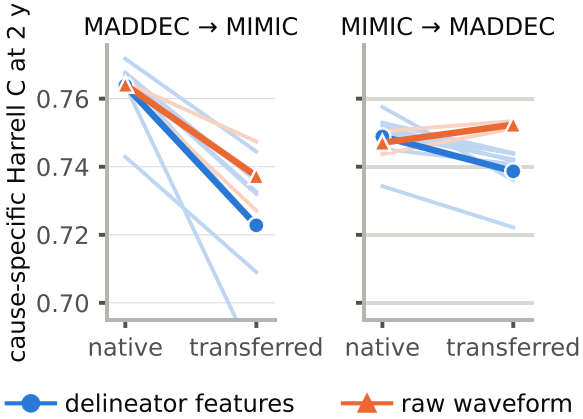

Figure 1. Effect of transfer learning in both ways. Faint lines are individual architectures, bold is the mean.

4. Discussion

In our study, we showed that architecture had a very limited effect on the output, about 0.02-0.03 C-index around the 0.75 area, while the choice of the negative set had much higher impact. Interestingly, while the proprietary 12SL won overall on point estimates on the best arm on both cohorts, the open delineator with 406 features was only marginally behind. The delineator is reliant on a deep learning architecture as well and not handcrafted features, such as the 12SL. An advantage for the waveform was present in transfer learning, where it transported better with a lower drop, and even slightly improved in one specific case, showing that it may generalize better even without any domain adaptation.

The results open up a discussion on what kind of feature extraction methodology should be used. A fiducial-point based set of features may be more useful for secondary analysis purposes, such as explainable AI. For an abstract representation analysis, such as latent space analysis, the raw waveform based might be better. The representations of the foundation model were seen to be as useful as the tabular models, and also on the same level as the CNN. For the predictive task, it did not seem to be that significant what architecture family was chosen for the task.

On identical hardware, the computing cost vastly differed between the architectures. One nested search was estimated to be between 0.21-0.38 hours on XGBoost, and 8.8-19.4 hours for the CNN. Search cost was uncorrelated with accuracy across cells.

4.1. Limitations

Both datasets were single hospital. No calibration results were reported. Only AF was analyzed for the cross-cohort analysis, as SCD was not available. The number of ECGs differed largely between the cohorts. The AF label used only the ECG-reading, which we believe is largely viable for MADDEC but not necessarily for MIMIC, where ICD code inclusion would have added 18.9% more AF events. Finally, the FM was pretrained on MIMIC and could not be fairly evaluated on it, and the Optuna trials for FM were only for its embeddings, which may understate its potential.

5. Conclusion

We studied the effect of the processing of an ECG in a predictive task with two endpoints, and three cohorts. We found that XGBoost had the best performance and was fastest to train, over or on par with a CNN and a foundation model. For predictive tasks on ECGs, it might be more worthwhile to focus on methodological aspects beyond the model architecture.

Acknowledgments

This work was funded by the European Union's Horizon Europe programme under Grant Agreement no. 101137278 (CVDLink). We thank Jad Haidamous for providing the ECG foundation model.

References

[1] K. C. Siontis, P. A. Noseworthy, Z. I. Attia, and P. A. Friedman, "Artificial intelligence-enhanced electrocardiography in cardiovascular disease management," *Nat. Rev. Cardiol.*, vol. 18, no. 7, pp. 465–478, Jul. 2021, doi: 10.1038/s41569-020-00503-2.

[2] S. Raghunath *et al.*, "Deep Neural Networks Can Predict New-Onset Atrial Fibrillation From the 12-Lead ECG and Help Identify Those at Risk of Atrial Fibrillation-Related Stroke," *Circulation*, vol. 143, no. 13, pp. 1287–1298, Mar. 2021, doi: 10.1161/CIRCULATIONAHA.120.047829.

[3] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, in NIPS '22. Red Hook, NY, USA: Curran Associates Inc., Nov. 2022, pp. 507–520.

[4] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Inf. Fusion*, vol. 81, pp. 84–90, May 2022, doi: 10.1016/j.inffus.2021.11.011.

[5] J. A. Hernesniemi *et al.*, "Extensive phenotype data and machine learning in prediction of mortality in acute coronary syndrome - the MADDEC study," *Ann. Med.*, vol. 51, no. 2, pp. 156–163, Mar. 2019, doi: 10.1080/07853890.2019.1596302.

[6] B. Gow *et al.*, "MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset." PhysioNet, 2023, version 1.0. doi: 10.13026/4NQG-SB35.

[7] A. E. W. Johnson *et al.*, "MIMIC-IV, a freely accessible electronic health record dataset," *Sci. Data*, vol. 10, no. 1, p. 1, Jan. 2023, doi: 10.1038/s41597-022-01899-x.

[8] N. Strodthoff *et al.*, "PTB-XL+, a comprehensive electrocardiographic feature dataset," *Sci. Data*, vol. 10, no. 1, p. 279, May 2023, doi: 10.1038/s41597-023-02153-8.

[9] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage AK USA: ACM, Jul. 2019, pp. 2623–2631. doi: 10.1145/3292500.3330701.

Address for correspondence:

Roni Ahola.
Tampere University, Faculty of Medicine and Health Technology
Korkeakoulunkatu 3, 33720 Tampere, Finland
roni.ahola@tuni.fi